SideIO: A Sided I/O System Framework for Hybrid Scientific Workflow

Dan Huang[1], Jiangling Yin[1], Jun Wang[1], Qing Liu[2]
[1]Department of Electrical Engineering and Computer Science
[1]University of Central Florida, Orlando, FL
[2]Oak Ridge National Lab, Oak Ridge, TN
[1]{jyin, dhuang, jwang}@eecs.ucf.edu [2]{liuq}@ornl.gov

***Abstract*—Recent years have seen an increasing number of *hybrid* scientific applications. These applications are characterized by two interwined components, an HPC simulation or experiments, and data analytics. Examples of such hybrid applications include fusion experiments such as those conducted at International Thermonuclear Experimental Reactor (ITER) and Joint European Torus (JET), and high-fidelity fusion simulations on high-end systems. Data analytics is a key aspect of many hybrid scientific applications. This process involves analyzing data from previous runs and using the knowledge gained to determine how the next experiment should be adjusted. This can lead to improved efficiency and a lowered cost of running experiments. Another hybrid application is Quantum Monte Carlo Package (QMCPack), which produces 250GB of analysis data every minute for production runs on 65536 compute cores. This sheer volume of data QMCPack produces a huge challenges to analytics.**

Unfortunately, current computing platform settings do not accommodate this emerging workflow very well. This is mainly because most HPC systems today employ Parallel File System (PFS) that is connected via high-speed networks to store data coming out of simulations. To perform analytics on data generated by simulation, data has to be migrated from storage to the compute nodes that are allocated to analytics. This data migration could introduce severe latency and energy consumption, given the ever-increasing data size involved with analytics.

While supercomputers equipped with PFS storage clusters still represent the mainstream HPC, many small-medium sized HPC clusters have been built to facilitate hybrid scientific workflow applications in fast-growing cloud computing infrastructures such as Amazon HPC cluster instances. In contrast to traditional supercomputer settings where there are high-speed links, the limited network bandwidth in scale-out HPC clusters implies that data migration will not be feasible. To address this problem, we propose a Sided I/O System Framework (SideIO) to avoid such migration overhead for small-medium sized HPC analysis clusters. Our main idea is to create a sided I/O path for HPC simulation programs to conduct analysis data to a staging cluster (small-medium sized HPC cluster equipped with DIFS), and than the analytic programs are performed on the staging cluster.

There are three contributions in SideIO. Firstly, an I/O separator will be designed to automatically extract and conduct analysis data into staging cluster equipped with DIFS (e.g. HDFS). Our preliminary experiments show that the throughput of N-to-N write on DIFS is 2 to 5 times lower than on PFS (i.e. Lustre). In light of this, in the second component, a unified write accelerator, run on the staging cluster to promote original HPC simulation programs writing scientific data into DIFS with comparable bandwidth to PFS. In order to reduce read-write I/O contention while running analytic programs on the same staging cluster with DIFS, an I/O scheduler and algorithm dynamically smoothe out both burst disk write and read traffic in DIFS for both simulation and analytic programs.

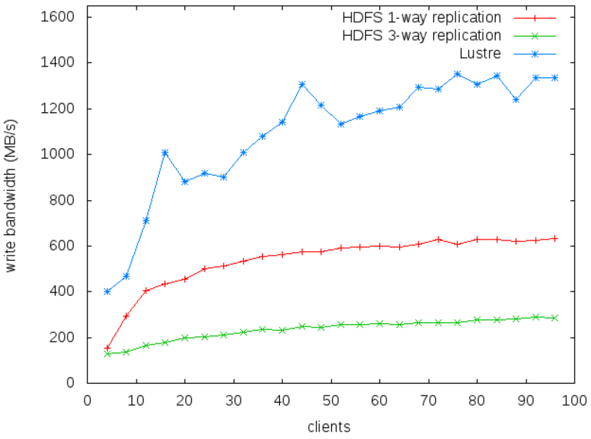

Figure 1: N-to-N Write Test on HDFS and Lustre

I. Appendix A

Our preliminary experiments are conducted on 61-node cluster testbed. Each node has dual 1.6GHz AMD Opteron processors, 16GB of memory, Gigabit Ethernet, and a 2TB Western Digital SATA disk drive. For our experiments, all nodes are connected to the same switch. On the testbed, MPICH is installed as a parallel programming framework on all compute nodes running CENTOS55-64 with kernel 2.6. The data intensive file system(HDFS) and parallel file system(Lustre) are both configured as follows: one node for the NameNode/JobTracker, 48 nodes as the DataNode/TaskTracker and other 12 nodes as client nodes. HDFS is configured as 3-way replication and 1-way replication respectively. The parallel I/O benchmark is MPI-IO Test, developed by Los Alamos National Lab. The experiment results are shown in the Figure 1.